Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 29 (2012) 1319 – 1324

Procedia Engineering

www.elsevier.com/locate/procedia

2012 International Workshop on Information and Electronics Engineering (IWIEE)

Semi-Supervised Locality Discriminant Projection

Yu Mao*, Yanquan Zhou, Hao Yu, Li Wei, Xiaojie Wang

Center of Information Science and Technology, Department of Computer Science ,Beijing University of Posts and Telecommunications ,10086 BeiJing, China

Abstract

In this paper, we consider the problem of semi-supervised dimensionality reduction. We focus on the local geometric structure of data and propose a novel method, called Semi-supervised Locality Discriminant Projections (SSLDP). It uses both labeled and unlabeled samples. Specifically, the labeled samples are used to explore the discriminating [a]information including both similarity and dissimilarity information, while the unlabeled samples are used to estimate the intrinsic geometric structure of data. Thus, SSLDP learns a discriminant projection which can best preserve both the discriminating structure and the local geometric structure of data. We evaluate SSLDP in the similarity measure which plays a key role in most of the information processing tasks. The experimental results show the effectiveness of our algorithm.

Keyword： dimensionality reduction, semi-supervised learning, manifold learning, Locality Preserving Indexing, Locality Discriminating Indexing；

1. Introduction

In many visual or text analysis applications, such as image or document retrieval, face recognition, text categorization, etc., one is often confronted with high dimensional data. It is considerable to represent the data in a lower dimensional space before performing efficient clustering or classification algorithms. So far, most methods of dimensionality reduction can be parted into two groups. The methods in the first group are derived from the global statistical properties of data, two famous techniques of which are Principle Component Analysis (PCA)[1] and Linear Discriminant Analysis (LDA)[2]. Specifically, PCA

* Corresponding author..
E-mail address: maomaoyu10@gmail.com.

doi:10.1016/j.proeng.2012.01.134

is an unsupervised method aiming at the optimal reconstruction of the data, and LDA is a supervised method which focuses on a more distinguishable projection. Both of them have been effectively employed in a variety of real-world applications, however, they only pay attention to the global statistical properties of data, and therefore often fail in the case where the data was highly nonlinear or the distribution of data was far away from the *Gaussian* distribution. To address this problem, many dimensionality reduction methods based on manifold learning have been derived, forming the second group. These methods focus on the intrinsic geometric structure of the data, and naturally can find more suggestive lower dimensional structure underlying the high dimensional observed samples. Three impressive algorithms of this group are Laplacian Eigenmap[3], Locality Preserving Projections[4] and Locality Discriminating Indexing [5].

In general, supervised methods are more efficient since they incorporate discriminating information. However, when given insufficient labeled data, the performance of supervised methods would be hardly guaranteed. It this case, taking the unlabeled data into account may greatly useful [6]. Hence, in this paper we consider the problem of semi-supervised dimensionality reduction based on manifold learning, expecting it to perform more effective and stably by taking both the labeled and unlabeled samples into account.

The rest of this paper is organized as follows. The Section 2 is devoted to a step-by-step introduction to the algorithm of Semi-supervised Locality Discriminant Projections (SSLDP). The experimental results are presented in Section 3. Finally, we conclude the paper and give suggestions for future work in Section 4.

2. Semi-supervised Locality Discriminant Projection

Before presenting the SSLDP, we firstly introduce the notations used throughout this paper. Let the matrix $X = \left[x_1, x_2, ..., x_N \right]$ $x_i \in \mathbb{R}^m$ denote the entire data set, where N is the number of samples and m is the feature dimension. $y_i = A^T x_i \in \mathbb{R}^d \left(d \le m \right)$ is low-dimensional representation of xi through linear transformations $A_{n \times m}$ .The complete data set $D = \left\{ \left(x_1, t_1 \right), \left(x_2, t_2 \right), ... \left(x_l, t_l \right), x_{l+1}, ... x_{l+u} \right\}$ (t_i is the label of x_i) is naturally parted into two subset: the labeled data set $L = \left\{ \left(x_1, t_1 \right), \left(x_2, t_2 \right), ... \left(x_l, t_l \right) \right\}$,and the unlabeled data set $U = \left\{ x_{l+1}, ... x_{l+u} \right\}$. In this case, the number of columns of X , i.e. N equals *l*+*u*.

2.1. Construct the adjacency graph

We use the adjacency graph to describe the intrinsic geometry of data. Choosing the appropriate type of graph and its parameters is not a trivial task. In general one should try to ensure that the local neighborhoods induced by this construction are “meaningful”. However, since there is no essential difference for our algorithm with various methods of graph building, we present our algorithm with k nearest neighbor graph.

Once the adjacency graph is constructed, we need to weight the edges between data points. We hope the weights of edges can reflect both of the natural geometric structure of data and the label information. Furthermore, we intend to involve both the similarity and dissimilarity information, since we know which data points have same label while which data points have different labels. Hence, we choose the special definition as following:

definition 1：

Wij = 1 if x_i and x_j are connected, and they are not in different classes

-1 if x_i and x_j are connected, and they are in different classes

0 otherwise.

Note that there are three cases that the edge between two connected data points should be weighted by 1. The first case is that the two data points are both labeled and their labels are the same; the second is that both of them are unlabeled, in which case we can't make sure about whether they are in different categories; the third is that either of them is unlabeled, in which case the label of the other could be neglected for the same reason mentioned in the second case. In fact one is usually encountered with the latter two cases since the labeled data are sparse.

2.2. Object Function and the Optimal Solution

We consider the following three sub-objective problems:

- We hope the samples that are in same class get closer in their local geometric structure, which can be formulated as:

$$\text{Min} \sum_{i,j} \left\| y_i - y_j \right\|^2 \quad (x_i, x_j \text{ are connected and in the same class}).$$

- We hope the samples that are in the different classes in the local geometric structure move further, which is formulated as:

$$\text{Max} \sum_{i,j} \left\| y_i - y_j \right\|^2 \quad (x_i,\ x_j \text{ are connected but in different categories}).$$

Note it equals to

$$\text{Min} -\sum_{i,j} \left\| y_i - y_j \right\|^2 \quad (x_i, x_j \text{ are connected but in different categories}).$$

- We have an additional purpose that if we cannot determine whether the two samples are in different categories, then, the mapped yi and yj should be "close" if xi and xj are "close". This intuitively leads to:

$$\text{Min} \sum_{i,j} \left\| y_i - y_j \right\|^2 \quad (x_i, x_j \text{ are connected and they are not in different classes}).$$

By considering the above three sub-objective problems together with the weight we have defined in **definition 1**, we present our objective function of SSLDP as following:

$$A^* = \arg\min_A \frac{1}{2} \sum_{i,j \in L} \left\| y_i - y_j \right\|^2 W_{ij}$$

which equals to:

$$A^* = \underset{\substack{YDY^T = I \\ YD1 = 0}}{\text{argmin}} \ \text{tr}\left(\mathrm{A^T X L X^T A} \right) \tag{1}$$

Where $Y = A^{\mathrm{T}} X$, D is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$. $L = D - W$ is the Laplacian matrix. The constraint $YDY^T = I$ removes an arbitrary scaling factor in the embedding, and the constraint $YD1 = 0$ eliminates the trivial solution that collapses all vertices of G onto the real number 1.

Solutions of Eq.(1) can be provided by the eigenvectors of the following generalized eigenvalue problem:

$$\mathrm{XLX^T a} = \lambda \mathrm{XDX^T a} \tag{2}$$

Let the column vectors $a_1, a_2, ..., a_d$ be the eigenvectors with respect to the first d non-zero minimum eigenvalues of Eq.(2) and let the transformation matrix $A_{n\times d} = [a_1, a_2, ..., a_d]$. A data point can be embedded into d dimensional subspace by following transformation:

$$x \to y = A^{\mathrm{T}} x$$

3. Experimental Results

In this section, several experiments were performed to evaluate our proposed algorithm. We apply the SSLDP algorithm to the task of similarity measure for the accuracy of similarity measure plays a key role in most of the information processing tasks including document clustering, classification, retrieval, etc. We also compared our algorithm with LPI[7] and LDI[5].

3.1.1. Data Preparation

We use Reuters-21578 as our data collection. Documents that appear in two or more categories were removed, leaving us 8293 documents consisting of semantic categories (topics). We kept the largest 20 categories which contain 7794 documents in total and the details are listed in Tabel 1. To compare with LPI, we use the keywords reported by [7] except the word "five" as it appeared in our stop-word list and was removed during the preprocessing. For each keyword *qi*, let D*i* denote the set of the documents that containing *qi*. Finally, we get 29 document subsets that contains multiple topics each. The sizes of these document subsets and the numbers of topics contained are listed in Tabel 2. Note that these subsets are not necessarily disjoint. We removed the stop words and no further preprocessing was done. For each subset D*i*, we use the first 10,000 words with highest frequency as features, and represent the documents as vectors in the resulted vector space using the Term Frequency (TF) indexing scheme.

Table 1. 20 semantic categories from Reuters-21578 used in our experiments.

category	**num of documents**	**category**	**num of documents**	**category**	**num of documents**
Earn	3713	sugar	114	alum	45
Acq	2055	coffee	110	grain	45
crude	321	gold	90	copper	44
trade	298	money-supply	87	jobs	42
money-fx	245	gnp	63	reserves	38
interest	197	cpi	60	rubber	38
Ship	142	cocoa	53		

Table 2. 20 The size of documents and topics related to the 29 keywords

Subsets	**num of documents**	**num of topics**	**Subsets**	**num of documents**	**num of topics**
agreement	761	16	Losses	237	17
American	350	15	Money	132	17
Bank	636	19	national	380	19
Control	235	14	Prices	587	19
Domestic	256	15	production	401	17
Export	263	17	Public	285	17
Exports	316	20	Rates	239	19
Foreign	442	19	Report	327	19

Growth	299	19	services	246	13
Income	346	13	Sources	255	15
Increase	581	19	Talks	284	15
industrial	238	15	Tax	549	14
Industry	392	18	Trade	594	20
international	686	20	World	359	19
investment	572	18			

3.1.2. Experimental Design

For each document subset D*i*, we compute its lower dimensional representations D*i*' by using SSLDP, LPI and LDI, then we evaluate the similarity measure between the documents in D*i*'. Intuitively, we expect that similarity should be higher for the document pair related to the same topic (intra-topic pair) than for the pair related to different topics (cross-topic pair). Therefore, we adopted the average precision used in TREC[8], regarding an intra-topic pair as a relevant document and the similarity value as the ranking score. Let *pi* denotes the document pair which has the *i*-th highest similarity value among all pairs in the subset D*i*. For each intra-topic pair *pk*, its precision is evaluated as follows:

$$precision(pk) = \#\text{ of } intra\text{-}topic \text{ pairs } pj \,/\, k,\ j \leqslant k$$

The average of the precision values over all intra-topic pairs in D*i* was computed as the average precision of D*i*.

For LPI, the topic information of documents are not used before valuation. For SSLDP and LDI, we use the topic of documents as their labels, in other words, the documents related to same topic is thought of belonging to same category. We refer the documents whose topic information are used as labeled data points while the documents whose topic information are neglected as unlabeled data points. To exam the performance of SSLDP with different size of labeled data points, we take varied proportion (from 1% to 50%) of documents related to each topic to compose labeled data set. Following the [7], for all algorithms, the number of nearest neighbors is set to be 7, and the cosine similarity is used.

3.1.3. Results

In Figure1, we compare the "best" and "mean" overall average precision of SSLDP, LPI and LDI with the number of dimension varies from 1 to 60. It can be seen that by incorporating the discriminating information, SSLDP outperforms LPI, and achieves at most 3.71% improvement when there is sufficient labeled data points. In the two sub-figures to the right, we can see that SSLDP performs much better than LDI. We note [5] uses LDI with 3 nearest neighbor graph, the value of parameter k here may not appropriate for this method. However, since this value is set without optimizing for SSLDP either (just following the [7]), we can argue that SSLDP is more robust to the changes of this parameter and more efficient in learning discrimant projection.

4. Conclusions

In this paper, we propose a new linear dimensionality reduction algorithm called Semi-supervised Locality Discriminant Projection (SSLDP). Specifically, the labeled data points are used to explore the discriminating information including both similarity and dissimilarity information, while the unlabeled data points are used to estimate the intrinsic geometric structure of data. Thus, SSLDP learns a discriminant projection which can best preserve both the discriminating structure and local geometric structure of data. We evaluate SSLDP in similarity measure on real-world dataset. The experimental

results demonstrate the effectiveness of our algorithm. In future work, we will explore the local and global consistency method for dimensionality reduction. We will also try different approaches to encode the dissimilarity information.

Acknowledgements

This paper was supported by Mechanism socialist method and higher intelligence theory of the national natural science fund projects (No. 60873001).

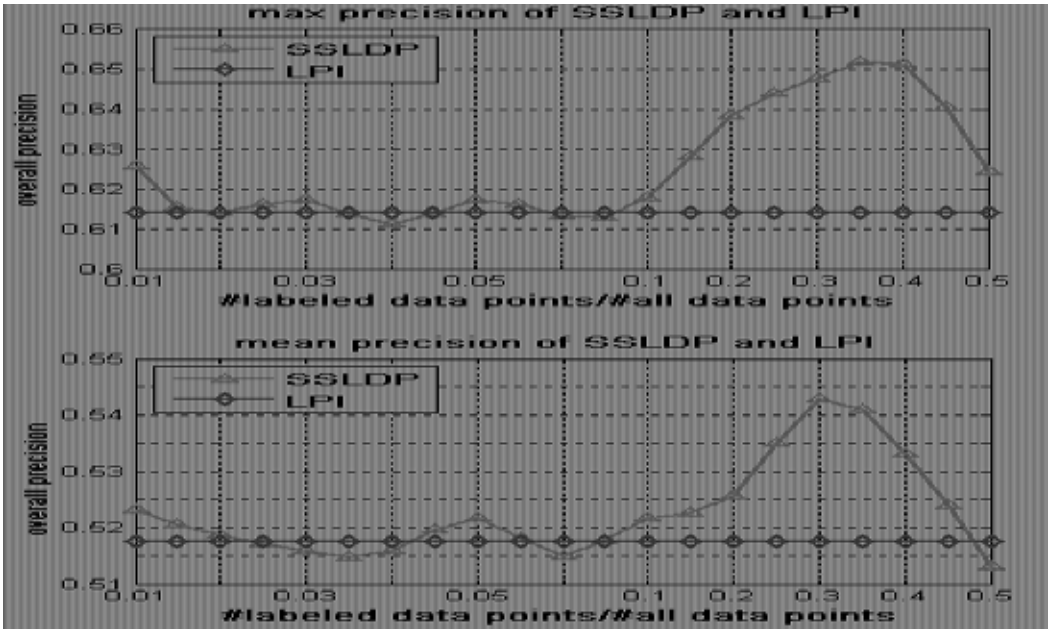

Fig. 1. The overall average precision of SSLDP with LPI and LDI with different size of labeled data points

References

[1] K. V. Mardia, J. T. Kent, and J.M. Bibby. Multivariate Analysis. Academic Press, 1980.

[2] K. Fukunaga. Introduction to Statistical Pattern Recognition. Academic Press, 2nd edition, 1990.

[3] Belkin M，Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 15,1373-1396, 2003.

[4] Xiaofei He and Partha Niyogi. Locality Preserving Projections. In Advances in Neural Information Processing Systems 16, NIPS'16, Vancouver, Canada, 2003.

[5]Jiani Hu, Weihong Deng, Jun Guo, and Weiran Xu. Locality Discriminating Indexing for Document Classification. SIGIR'07, Amsterdam, The Netherlands. , 2007, ACM 978-1-59593-597-7/07/0007.

[6]Deng Cai, Xiaofei He, Jiawei Han. Semi-supervised Discriminant Analysis. Eleventh IEEE International Conference on Computer Vision, ICCV 2007.

[7]Xiaofei He, Deng Cai, Haifeng Liu, Wei-Ying Ma. Locality Preserving Indexing for Document Representation. SIGIR'04, July 25–29, Sheffield, South Yorkshire, UK. , 2004

[8]R. K. Ando. Latent Semantic Space: Iterative Scaling improves precision of inter-document similarity measurement. In Proc. of the 23th International ACM SIGIR, Athens, Greece, 2000.